Mentalizing During Social InterAction: A Four Component Model

Haiyan Wu[1,2,3], Xun Liu[1,2], Cindy C. Hagan*[3] Dean Mobbs*[3,4]

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology

[2] Department of Psychology, University of Chinese Academy of Sciences;

[3] Division of Humanities and Social Sciences, California Institute of Technology

[4]Computation and Neural Systems Program at the California Institute of Technology

Correspondence: chagan@caltech.edu or dmobbs@caltech.edu

Abstract

Mentalizing, conventionally defined as the process in which we infer the inner thoughts and intentions of others, is a fundamental component of human social cognition. Yet its role, and the nuanced layers involved, in real world social interaction are rarely discussed. To account for this lack of theory, we propose the Interactive Mentalizing Theory (IMT). We discuss the connection between mentalizing, metacognition, and social interaction in the context of four elements of mentalizing: (i) Metacognition - inference of our own thought processes and social cognitions and which is central to all other components of mentalizing including: (ii) first-order mentalizing – inferring the thoughts and intentions of an agent's mind; (iii) personal second-order mentalizing - inference of other's mentalizing of one's own mind; (iv) Collective mentalizing: which takes at least two forms (a) vicarious mentalizing: adopting another's mentalizing of an agent (i.e. what we think others think of an agent) and (b) co-mentalizing: mentalizing about an agent in conjunction with others' mentalizing of that agent (i.e. conforming to others beliefs about another agent's internal states). The weights of these four elements is determined by metacognitive insight and confidence in one's own or another's mentalizing ability, yielding a dynamic interaction between these circuits. To advance our knowledge on mentalizing during live social interaction, we identify how these subprocesses can be organized by different target agents and facilitated by combining computational modeling and interactive brain approaches.

Keywords: Metacognition, mentalizing, vicarious mentalizing, co-mentalizing, social inference

1. Introduction

Niccolo Machaivelli was, perhaps, the first to introduce the concept of understanding another's mind, stating in his 1513 book *The Prince* "Minds are of three kinds: one is capable of thinking for itself; another is able to understand the thinking of others; and a third can neither think for itself nor understand the thinking of others." Three centuries later, Charles Darwin spoke of his father's remarkable ability to read the character of others and "even the thoughts of those whom he saw even for a short period of time." Machaivelli and Darwin's prescient insights capture the idea that we have evolved a system to gain access to others' inner thoughts. In modern psychology, this system is referred to as mentalizing or the process by which we infer the content of the thoughts of others (Baron-Cohen, Leslie, & Frith, 1985; Frith & Frith, 1999). The process of mentalizing occurs when we believe that something has a mind (Gray, Gray, & Wegner, 2007) or when we project a mind into a mindless object (e.g. anthropomorphism). Whether we are right or wrong, mentalizing is the process of theorizing about a target agent's inner beliefs and allows us to imbue goals and expectations onto others.

Dynamic aspects of social information are often combined to infer another agent's belief or attitude, either from their tone of voice, facial expressions, body postures, or some combination thereof (Bögels, Barr, Garrod, & Kessler, 2014; Gallagher & Frith, 2003; Hagan et al., 2013). To understand another person's thought processes, the agent needs to retrieve, enquire about, or infer their experiences. Together these provide the context that enables insight into the encoder's mind. The process by which this occurs is critical as it refines the receiver's interpretation of social signals, leading to a better understanding of others, which is vital for successful social interaction. This understanding is generated through perspective taking, simulations, active learning, and predictions that are dynamically updated (Silston, Basset and Mobbs, 2018).

Here, we propose an Interactive Mentalizing Theory (IMT), where mentalizing is divided into the following four components: 1) metacognition or insights and confidence about one's own thought processes and cognition. This sets the foundation to all other mentalizing processes. 2) first-order mentalizing - mentalizing about an agent's mind 3) personal second-order mentalizing - inference of other's mentalizing of one's own mind, and 4) collective mentalizing: vicarious mentalizing adopting another's mentalizing of minds of other agents, and co-mentalizing or weighted second-order mentalizing, which usually occurs when we believe that others have better insights than do we into the thoughts of relevant agents. Metacognitive insight and confidence about one's own mentalizing ability determines the weights ascribed to all four elements of the IMT. These four elements of the IMT have overlapping neuronal ensembles, where population codes integrate to form a coherent picture of others' thoughts and intentions.

2. Metacognition and Mentalizing

A key element of our model is that the aforementioned components of mentalizing are influenced by metacognitive processes. Metacognition refers to second order cognitions including thoughts, knowledge, or beliefs about one's thoughts or thought processes (Flavell, 1979; Metcalfe & Shimamura, 1994; Nelson, 1990). Broadly speaking, metacognition includes knowledge or self-awareness about beliefs, mental-states, motivations, intentions and abilities, as well as the ability to consciously and deliberately monitor and regulate these aspects (Flavell, 1979). Various aspects of metacognition include meta-level abilities that evaluate and control different cognitive processes. One noteworthy bias is that people often believe themselves to be less susceptible to biasing influences than others (Pronin & Kugler, 2007).

Although it's not easy to quantify metacognitive sensitivity or metacognitive accuracy, attempts have been made to do this (Fleming & Lau, 2014). For example, studies have

utilized signal detection theory to compute a "response-specific" meta-d′ and ROC Curves to measure metacognitive sensibility (Fleming, Weil, Nagy, Dolan, & Rees, 2010; Maniscalco & Lau, 2014). A reliable index is metacognitive accuracy, the correlation between subjective confidence and the actual performance. Indeed, subjective levels of confidence is a common metric for assessing metacognitive estimates of accuracy (Koriat & Goldsmith, 1996). As one of the most essential dimensions of metacognition, confidence refers to people's subjective beliefs about the validity of their thoughts and judgments, irrespective of objective accuracy measures.

During social interaction, an online metacognitive process takes place at a first-order level, second-order level, or group/collective level (Bahrami et al., 2012; MMahmoodi et al., 2013). Social interaction influences the way in which these metacognitive processes are updated or confirmed, either through experiences such as social acceptance or rejection, or social feedback or persuasion (Petty, Briñol, &Tormala, 2002). Metacognition also regulates the subjective social experience, during social learning and social interaction. The representation of one's own social competence might drive others to perceive him/her in a similar vein, encouraging an alignment of behavior. For instance, levels of confidence projected during an interactive negotiation or business transaction may largely determine the economic gain (Charness, Rustichini, & van de Ven, 2018).

Throughout this article, we will focus on metacognition as part of the 'global mentalizing system', which we believe plays a major role in social inference in interaction. Over the past few decades, there was a gap between metacognition and mentalizing, as metacognition mainly pertains to one's own mind and mentalizing involves inferences about another agent's mind. Theoretical models that link metacognition with mentalizing propose that mentalizing is a knowledge-oriented dimension of metacognition (Efklides & Misailidi, 2010; Kuhn, 2000). Different aspects of social interaction presumably involve metacognition: 1) social

judgment (Petty, Briñol, Tormala, & Wegener, 2007), for example, metacognition in attitude change (the confidence in old and new attitudes after exposure to persuasive information) (Tormala & Petty, 2004; Rollwage, Dolan & Fleming, 2018); 2) representation of the social knowledge or beliefs of others (see review Frith, 2012), 3) together with mentalizing, switching the i-mode and we-mode for joint action (Tuomela, 2006) or collective decisions with confidence sharing between individuals (Bahrami et al., 2010; Mahmoodi et al., 2013; Marshall, Brown, & Radford, 2017; Stasser & Titus, 1985). The subjective confidence with which we assign a correct internal estimate to our mental state inferences is crucial in interaction, although not always explicitly so. However, research lacks a theoretical framework that combines metacognition with mentalizing during social interaction.

3. Mentalizing

Mentalizing is the mental ability to understand other people's behavior in terms of their intentions, beliefs, needs, desires, or goals (Fonagy & Luyten, 2009). Mentalizing originally refers to the attribution of mental states to both others and oneself (Kidd & Castano, 2013; Premack & Woodruff, 1978; Zunshine, 2006), and with multiple dimensions such as cognitive and emotional dimensions (Choi-Kain & Gunderson, 2008; Fonagy & Luyten, 2009). Substantial literature shows that by nature, humans engage in mentalizing during tactical activities such as deception and lie detection (Granhag & Hartwig, 2008), and persuasion (Slaughter, Peterson, & Moore, 2013). Humans also engage in mentalizing during practical activities such as teaching and learning (Wang, 2015). Impairments in mentalizing ability has been reported in various neuropsychiatric disorders of development, such as autism (Baron-Cohen, Leslie, & Frith, 1986), schizophrenia (see review Brüne, 2005), and depressive disorders (Berecz, Tényi, & Herold, 2016).

Complex social interactions are cognitively demanding because of the depth or higher-order reasoning they require (Hampton, Bossaerts, & O'Doherty, 2008). Usually the zeroth-order of

mentalizing involves knowledge of some objective facts without necessitating a reasoning component; the first-order of mentalizing is the ability to understand that another person can have thoughts that are different to our own (e.g. "false" beliefs); the second order of mentalizing is the ability to understand that two individuals can have different thoughts, one of which is in alignment with our own thoughts and one of which is not; advanced mentalizing occurs when one is able to infer what one person thinks that another person is thinking (Banerjee, 2002). While these mentalizing abilities are thought to develop in individuals in an orderly fashion, the entire process occurs at different rates for different individuals, and is often completed in adolescence (Banerjee, 2002). We refer to mentalizing as an imaginative mental activity that involves at least two agents (self and other, or other and other) and two aspects (cognitive and affective) (Choi-Kain & Gunderson, 2008; Fonagy & Luyten, 2009).

4. Personal second-order mentalizing

As one aspect of higher-order mentalizing, personal second-order mentalizing is a novel component that we propose refers to one's metacognitive insight or confidence about others' mentalizing of one's inner thoughts to reveal how much insight we think that others have into our own inner thoughts and intentions (i.e. does he or she know what I'm thinking?). As people's thoughts can be oriented to external or internal stimuli including other people, the environment, or themselves, personal second-order mentalizing plays a crucial role in social interaction. We are specifically interested in mentalizing that involves both metacognitive and interpersonal components in social interactions, in which high-level metacognitive processes such as confidence, partially depend on the mentalizing processes involved in social interactions with others. This higher-order process may or may not be associated with reflective awareness, but it is nevertheless a metacognitive process. We believe that people can have varying degrees of subjective feelings of confidence about others' insights to one's own thoughts or beliefs. Personal second-order mentalizing holds that people's social

behavior not only relies on subjective mentalizing abilities but also on their awareness and confidence in their own ability to hide their inner motives during social interaction.

Personal second-order mentalizing is generated through adopting the perspective of another agent, simulating their mind, predicting their behaviors, and adjusting one's confidence based upon how well our prediction matched the observed outcome. For example, in an online interactive dishonesty study, the proposers thought that their opponents had poor insight into their own decisions, reported with higher confidence in the task, while responders with the higher-confidence in hiding their thought reported lower trust to the proposers (Wu, Fung and Mobbs, 2019). Further, the personal second-order mentalizing proposed in our theory is more about assessing one's own abilities -- it is not aimed at objectively discriminating between those with poor and good mentalizing abilities. Following from this notion, a related effect of metacognitive bias is laid out by Dunning-Kruger (Kruger & Dunning, 1999), and indicates that individuals may overestimate their own abilities relative to others (Kruger & Dunning, 1999; Dunning, 2011; Taylor & Brown, 1988). Taken together, self-evaluation may be biased such that we assign a higher rating to one's own mentalizing capacity, and a lower rating with respect to others' mentalizing capacities.

5.Collective mentalizing: Vicarious mentalizing and co-mentalizing

Collective mentalizing is when additional agents influence our inferences of a target agent's inner mental states. Vicarious mentalizing is when one conforms to other's views of an agent. This is different from co-mentalizing, which takes into consideration the perceived validity of the opinion of the one providing the information. However, vicarious mentalizing and co-mentalizing both manifest from an interaction with third-party others. Agents tend to adopt their peers' or the majority opinion (e.g., beliefs, choices and attitudes) during social interaction and align their beliefs or decisions under social influence (i.e., social conformity) (Cialdini & Goldstein, 2004; Wood, 2000). Conformity also depends on power dynamics and

hierarchies, as one has a tendency to conform to the beliefs and opinions of higher-ranking others (Qi, Footer, Camerer, & Mobbs, 2018). Conformity impacts many aspects of cognition. For example, people show memory conformity as they conform to another person's memory (Thorley, 2013), and a subsequent study (Wheeler, Allan, Tsivilis, Martin, & Gabbert, 2013) showed that explicit mentalizing simulation biases memory conformity toward the similar pre-labeled targets. Since social conformity affects many aspects of social cognition, we believe that it also plays a key role in mentalizing during social interaction.

Vicarious mentalizing occurs when we gain insight into a targeted agent's mind (e.g., A) using another agent's (e.g., B) insight into the targeted agent's mind. Such a component is critical when we need to rely on the opinion of others. For example, if we barely know Agent A, but we know Agent B and we know that Agent B has more experience with Agent A, we can summate that Agent B will have better insights into Agent A's thoughts and intentions. Vicarious mentalizing, therefore interacts with the metacognition (see Figure 1). Various factors such as the distance or relative position between the self and the target agent (A), the intermediary agent (B), adjacency or availability, connectedness, reputation etc., can affect this component. This may lead to another factor about the capacity to manipulate other's inference towards oneself or another agent while navigating the social world, especially in social contexts spanning more than two people. People take longer to perform vicarious mentalizing as people may experience both egocentric and allocentric interference on mentalizing, as the ability to take the perspective of the others is affected by the perspective of oneself. Vicarious mentalizing can lead to emotional arousal, preference, and attitude changes. For example, if we think Agent B dislikes agent A, we might also believe that there is something bad about agent A (Figure 1).

Co-mentalizing is built on the notion that humans are motivated to share their inner states (beliefs, mental) about and understanding of the social world (see Hardin & Higgins, 1996),

which in this case, refers to the concept of co-mentalizing. Co-mentalizing can be best conceptualized as co-operative or shared mentalizing processes with one another. For instance, when professors interview a new graduate student candidate, they tend to form their impressions jointly with their colleagues. The convergence of mentalizing leads to more confidence when people take into account the (inferred) inner states of others, especially significant others. Taking the perspective of others, adjusting the communication to a mutual understanding, or "shared reality" (Echterhoff, Higgins, & Levine, 2009), can form and maintain social bonds (Hardin and Higgins, 1996). According to the affiliative social tuning hypothesis in shared reality, in order to achieve a shared understanding, one is to "tune" their views toward the views of the other, with an overarching goal of affiliative motivation toward this person. Evidence indicates that humans need to share their inner states (beliefs, goals, thoughts, and feelings) and build a shared mind, or common goal, from infancy (see review from Liszkowski, 2018) to adults. In studies of mothering, Hrdy (2009) emphasizes how attachment and co-operation underpin the evolution of mentalization.

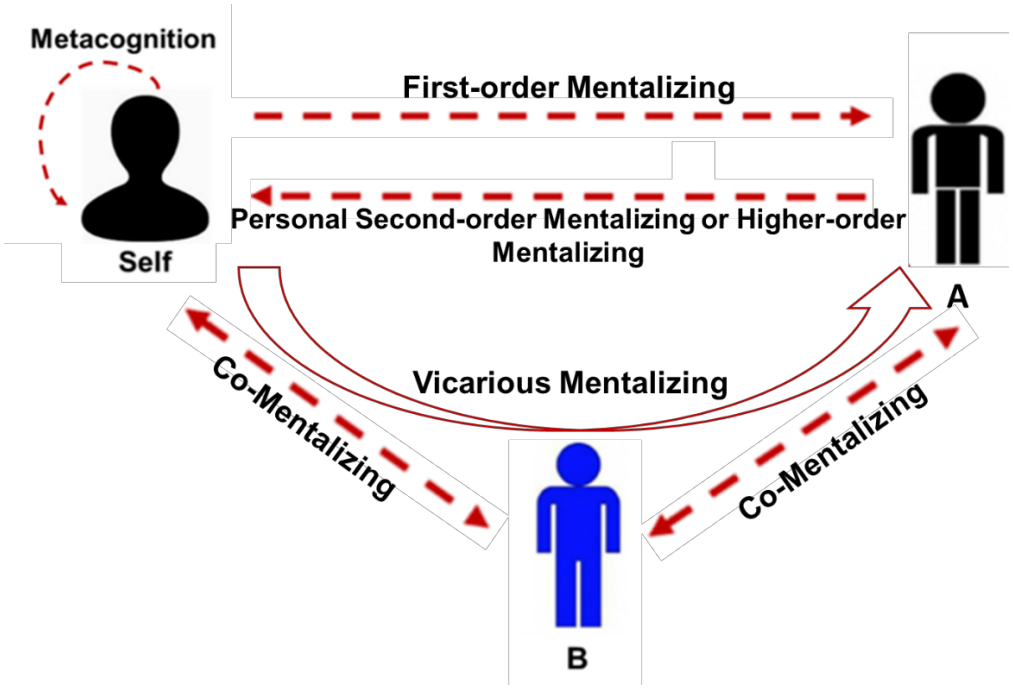

Figure 1. An depiction of the four elements in our Interactive Mentalizing Theory: 1) First-order mentalizing: mentalizing of others' mental states from the perspective of the self (i.e., self-other); 2) personal second-order mentalizing/higher-order mentalizing: mentalizing

of how much others can mentalize one's mental states from the perspective of others (i.e., other-self); and 3) vicarious mentalizing: mentalizing of other's mind from the perspective of others (i.e., other-other); and, 4) co-mentalizing: mentalizing of another's mind through others in combination with oneself. Metacognitive processes govern the weights we assign to the components within each element. In this figure, the self, A, and B, all could be the mentalizer, and people can mentalize other's thoughts in a direct or vicarious/indirect way. For example, the self can infer A's belief by mentalizing A directly or try to infer B's mentalizing about A.

6. The Interactive Mentalizing Theory (IMT)

We propose that four key inferential processes are evoked during social interaction: metacognition, first-order mentalizing, personal second-order mentalizing and collective mentalizing. The weight of these processes are contextually determined and may rely on internal monitoring. Our goal is to describe the multi-layered nature of human mentalizing with the influence of metacognitive insight, and enhance our further understanding of social navigation through vicarious mentalizing and co-mentalizing skills. For example, during social interaction, Agent A has inferences about Agent B's intentions. Agent A also has beliefs about how much insight Agent B has into their own internal states. Agent A also has metacognitive insight into their own thought and this is interpreted with high or low confidence. If the confidence is low, Agent A will look to others (e.g. Agent C) to guide their inferences of Agent B. The weight of these four processes alters across contexts and how much experience they have with an agent.

6.1. Relationship between mentalizing and personal second order-mentalizing

During real time social interaction, how we infer other mental states is impacted by how much we think they have insight into ours. In general, mentalizing and personal second order-mentalizing should be correlated but dissociable, with the meta-level monitoring and controlling of these two components dynamically. According to simulationists, our mind-reading capacity is an ability to project ourselves based on others' perspectives, by simulating their mental states with our own. Following this view, mentalizing capacity would be correlated with the personal second order-mentalizing component. The mentalizing ability, however, may not be acquired solely through simulation, and may involve other abilities such

as counter-factual thinking, social reasoning, and social learning (see Figure 2). Furthermore, due to the Dunning-Kruger effect, mentalizing would be negatively correlated with the personal second order-mentalizing component, as people who overestimate one's own ability may underestimate others.

6.2. Mentalizing and second order-mentalizing are influenced by metacognition

In our social inferential system, metacognition can be the core ability that links people's understanding about others (Gallese & Goldman, 1998). This view is partly supported by evidence that better self-reflection correlates with better understanding of others (Dimaggio, Lysaker, Carcione, Nicolo, & Semerari, 2008). People can have varying degrees of confidence or certainty about what others are thinking or feeling, but can generate beliefs and inferences. Mentalizing is viewed as a tool to predict others' behaviors and adjust estimates of others' mental states. Similarly, people rarely know others' thoughts about themselves without social interaction (low confidence or high uncertainty), but they infer these from dynamic social interactions. In this process, people's inferences originate from a self-generated first impression or feeling, and are updated with metacognitive thinking. Inferences that lead to a successful interaction will be viewed as a "correct" answer, or a reward signal.

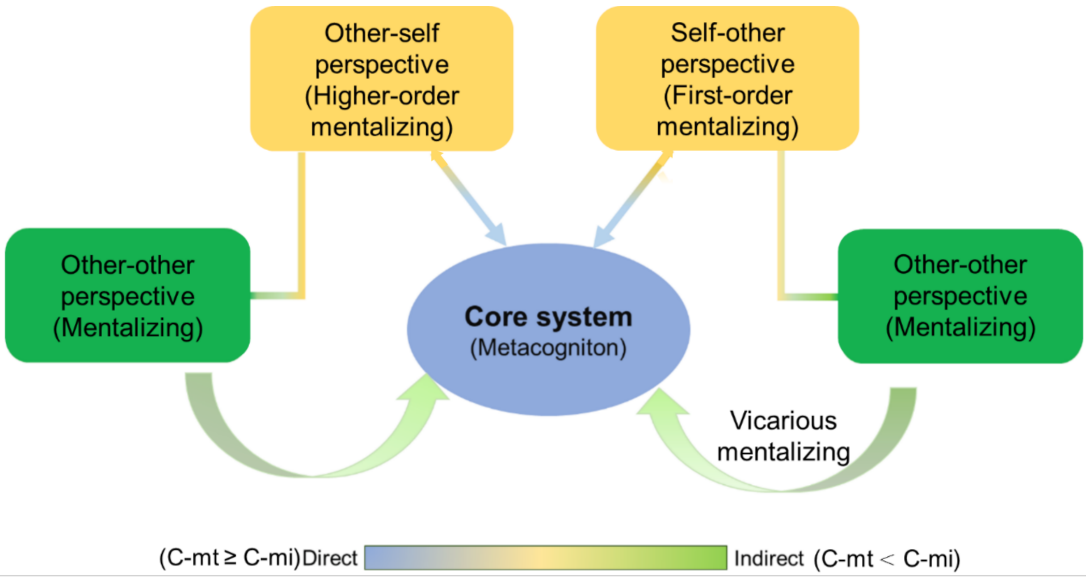

Figure 2. The box diagram of different inputs of mentalizing in social interaction. The core metacognition system receives inputs from one's mental states from the self-other perspective, other-other perspective (e.g., inferring other's mentalization about another agent) and other-self perspective; 2) People rely more on the other-other perspective (Vicarious mentalizing, or co-mentalizing) when the inputs from self-other perspective and other-self perspective are insufficient or with low self-confidence. C-mt: confidence about the mentalizing of the target agent, C-mi: confidence of the mentalizing to the intermediary agent (C-mi).

6.3. Low metacognitive confidence facilitates collective mentalizing

Within collective mentalizing, one needs to consider the personal level of confidence about their ability to mentalize two sources -- the target agent (C-mt) and the intermediary agent (C-mi). Necessarily, metacognitive confidence will vary upon relational distance, status, in/out groups, and adjacency among different agents. For example, one may prioritize inferring the target agent's mind directly if the (C-mt) is larger or equal to the (C-mi), or the distance to the both are the same, whereas an indirect/vicarious mentalizing path may be taken if direct path is unavailable, or the distance with the target agent is longer. A low level of metacognitive confidence will lead one to search for more information and may prompt one to conform to what we believe others think about an agent. e.g. I think Agent A has bad intention because I don't know Agent A but Agent B does and Agent B doesn't trust Agent A.

7. A neural model of the Interactive Mentalizing Theory

Numerous brain regions are known to be involved in mentalizing, and the brain network of different components show both areas of convergence and divergence (Figure 3). The medial prefrontal cortex (mPFC) is a central node in the social brain network and mentalizing, and also plays a key role in self-referential processing (Mitchell, Banaji, & Macrae, 2005). Moreover, mPFC is involved in metacognition, specifically in encoding decision confidence

(Bang & Fleming, 2018). Posterior superior temporal sulcus/temporoparietal junction (pSTS/TPJ) is another key brain region involved in mentalizing (Blakemore, 2008; Frith & Frith, 2001; Saxe & Kanwisher, 2003), with research supported by brain lesion studies (Apperly, Samson, Chiavarino, & Humphreys, 2004). The pSTS is involved in social perception (Pelphrey, Morris, & McCarthy, 2004; E. Redcay, 2008; Hagan et al., 2009; Hagan et al., 2013) and social attribution (Gao, Scholl, & McCarthy, 2012; Lee, Gao, & McCarthy, 2014; Schultz, Imamizu, Kawato, & Frith, 2004). Additionally, the pSTS and temporal lobe more generally may be viewed as a hub for integrating and processing signals of emotion for the purpose of mentalizing (Hagan et al., 2009; Hagan et al., 2013; Olson, Ploaker, & Ezzyat, 2007).

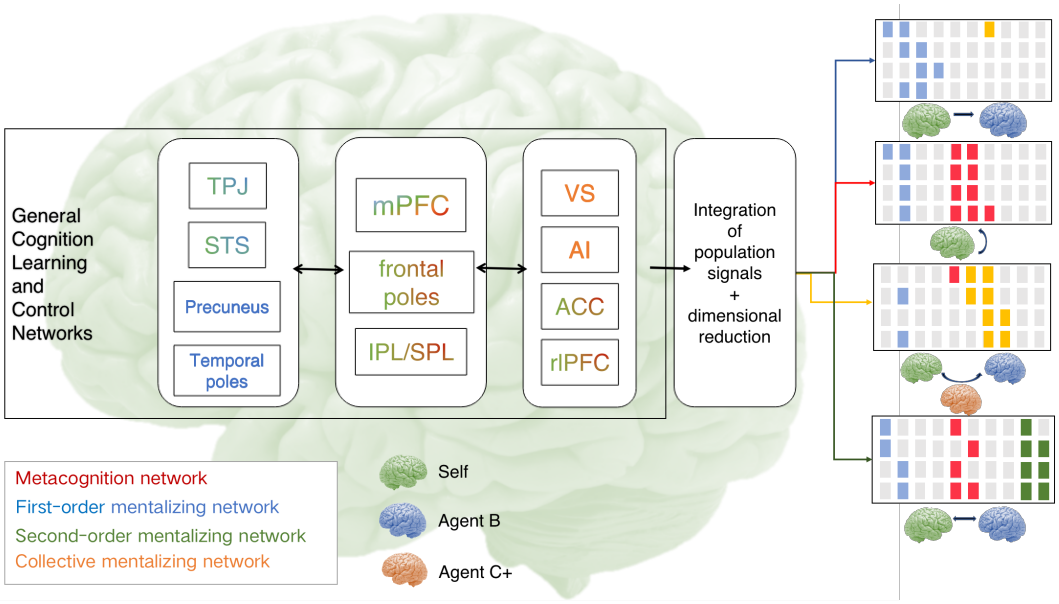

Figure 3. Simplified model of the different neural networks involved in mentalizing during social interaction. 1) The core metacognitive neural system receives inputs from both first-order and higher-order mentalizing system inputs; 2) anterior cingulate cortex (ACC) and lateral PFC (lPFC) is a junction for both self and other experience/inference, and also encodes prediction errors signal in social learning; 3) metacognitive neural system (mPFC, SPL and reward learning brain area like VS and AI) involved in collective mentalizing; 4) when people share own mentalizing with others and seek other's mentalizing, the processes can activate all of the four brain systems. The population of cells in these regions create a dynamic social wide brain network in different cases of mentalizing (see the right panel, active neurons are colored, while inactive ones are gray).

Mentalizing produces activity in a system of brain regions, including dorsomedial prefrontal cortex (dmPFC), temporoparietal junction (TPJ), posterior superior temporal sulcus (pSTS), and posterior cingulate cortex (PCC) which is more broadly involved in projecting one's self outside of the present moment and location (Buckner and Carroll, 2007, Spreng et al., 2009, Zaki and Ochsner, 2012). To the extent that vicarious, but not personal, reward involves mentalizing, these regions might be engaged preferentially by vicarious reward.

The frontal pole cortex (FPC) is involved in both metacognition via the self−evaluation of performance based on feedback and mentalizing (see meta-analysis from Gilbert et al. 2006). Evidence indicates that both metacognitive and mentalizing functions are associated with the lateral frontal pole activity. The lateral PFC (lPFC) and anterior cingulate cortex (ACC) are activated in retrospective judgments of confidence, which may link to the performance monitoring function (Fleming & Dolan, 2012). Apart from PFC, parietal cortex is also involved in subjective confidence. fMRI evidence indicate stronger intraparietal sulcus (IPS) or superior parietal lobule (SPL) activity for decisions associated with lower metacognitive confidence than higher metacognitive confidence (Chua, Schacter, & Sperling, 2009; Hongkeun & Roberto, 2007; Chua, Schacter, Rand-Giovannetti, & Sperling, 2006; Kim & Cabeza, 2009).

In our brain model of IMT, the lPFC and ACC receive input regarding the mental states of self and others. With access to metacognition, these two meta-level regions monitor and regulate the confidence about mental states. We believe that ACC is a potentially important region that integrates distinct forms of social information from various social networks such as social perception, motivation, and social prediction. For instance, ACC supports the detection of matched or mismatched signals between the prediction of another's intention in a competitive game (Gallagher, Jack, Roepstorff, & Frith, 2002). With the signals from dorsolateral prefrontal cortex (dlPFC) to other mentalizing brain regions, people adaptively

adjust their inferences (the complementary function in metacognition). According to the "simulation" and "projection" account, we believe that first-order mentalizing and personal second-order mentalizing networks closely interact with each other. The personal second-order mentalizing component, gets inputs from the first-order mentalizing network, and gives output. The medial frontal cortex in the personal second-order mentalizing network is more involved in self-referential thinking, while the lateral frontal cortex is involved in higher-order and strategic thinking. It is noteworthy that different kinds of mentalizing recruit distinct brain regions and networks. For instance, STS is primarily involved when people infer intentions from movement as it plays a role in biological motion and changeable cue processing (Castelli et al., 2000; Schultz et al, 2004; Saxe et al., 2004; Gobbini et al., 2007; Haxby et al., 2002), while vmPFC is suggested to activate more strongly during emotional mentalizing (Atique et al., 2011). There is also stronger functional connectivity between vmPFC and the left and right TPJ, than intention mentalizing (Atique et al., 2011).

Vicarious sharing of others' mentalizing often requires an understanding of how other's value a particular agent, especially when observers' and social targets' opinions diverge. As people need to vicariously learn from others (Bandura et al., 1963) and, sometimes, conform to another's belief, we highlight two brain networks involved in vicarious mentalizing, the vicarious learning network and the conformity network. The vicarious mentalizing brain largely overlaps with the belief prediction-errors brain (VS and AI) as a function of vicarious reinforcement, or observational learning (Burke et al., 2010). ACC plays a key role in encoding multiple types of prediction error signals and vicarious reinforcement learning (Lockwood et al., 2015; Apps et al., 2015, Hill et al., 2016). For example, ACC is involved in the learning of fear cues through observation (Olsson, Nearing, & Phelps, 2007).

As for the conformity component, considerable efforts have been made to unveil its neural mechanisms (Berns, Capra, Moore, & Noussair, 2010; Edelson, Sharot, Dolan, & Dudai,

2011; Izuma & Adolphs, 2013; Izuma et al., 2010; see Stallen & Sanfey, 2015; Toelch & Dolan, 2015, for a review). The majority of studies use functional magnetic resonance imaging (fMRI) and experimental paradigms in which participants were exposed to stimuli accompanied with another's choices (Bikhchandani, Hirshleifer, & Welch, 1998), judgments (Berns et al., 2005), ratings (Nook & Zaki, 2015) or advice (Biele, Rieskamp, Krugel, & Heekeren, 2011; Qi et al., 20 18). A meta-analysis of conformity literature showed that pMFC activity can predict the behavioral change (conformity) (Wu, Luo, & Feng, 2016). It is interesting that the pMFC is also involved in metacognitive judgments associates with lower confidence (see Figure 3 in Molenberghs, Trautwein, Böckler, Singer, & Kanske, 2016). Together, the evidence suggests that the key role of pMFC in the dynamics of mentalizing includes conformational mentalizing when one's confidence is low.

8. The Interactive Mentalizing Theory: implications and future directions

Our theory posits that in addition to a classic mentalizing system involved in self-other interaction, a metacognitive system oversees different types of mentalizing, leading to various possible interaction outcomes. Future research can support our theory or test our hypotheses through developing tools, testing situations involving social interaction or utilizing different demographic samples for evidence of conservation/replication, and computational modeling etc. We expect our theory can deepen our understanding of mentalizing deficits in clinical samples and will assist in the development of interventions by suggesting improvements in both metacognitive and mentalizing capacity.

8.1. The assessment of the components in Interactive Mentalizing Theory.

One concern about studying the mentalizing components is the methodology, as people's inner states are not observable. An important question therefore is whether the components in the mentalizing are indeed measurable. If so, are they then meaningful, and is it feasible to predict people's behavior? In our theoretical framework, the IMT model does not embrace all

mentalizing subprocesses. Rather, we aim to capture several aspects of mentalizing in social interaction, and thus focus on only four components (mentalizing, metacognition, and second-order mentalizing, collective mentalizing) (see Figure 1). We believe that these constructs are fundamentally related, and shaped heavily by social interaction, but independently measurable (e.g. Wu, Fung and Mobbs, 2019).

8.2. Implication of diverse approaches in interactive mentalizing

With the development of social decision neuroscience, an increasing number of studies investigate mentalizing during social interactions, with computational models and tasks adapted from game theory. For example, Yoshida et al (2008) adopted ideas from optimal control and game theory and provided a computational model for "game theory of mind," using model validation and updating techniques (Yoshida et al., 2008). One group has also proposed a k-ToM model, which predicts that the performance of agents engaged in competitive repeated interactions increases with their ToM sophistication (Devaine et al., 2014). We call for more efforts in computational models of different dynamic mentalizing components within the IMT framework. In order to map the network involved in dynamic mentalizing, we also argue for a greater focus on how different brain networks change dynamically (Silston, Basset and Mobbs, 2018) given a two-person context, third-person context or even group context (Redcay & Schilbach, 2019; Camerer and Mobbs, 2017).

8.3. Clinical implications of our mentalizing theory

Research suggests deficits in mentalizing capacity is one of the core features of personality disorders and clinical conditions such as autism, psychopathy, and may perhaps underlie the emotion labeling deficits present in broad-ranging clinical conditions such as alexithymia (Liotti & Gilbert, 2011; Lombardo, Chakrabarti, Bullmore, Baron-Cohen, & Consortium, 2011; McDonald & Flanagan, 2004). Children with autism are unable to understand others' beliefs in the false belief task, and such mentalizing failures are suggested to lead to impaired

social communications (Baron−Cohen et al., 1986). More interestingly, the performance of children with autism in a deception task is consistent with that of a false belief task, indicating that mentalizing deficits may impair complex social interaction abilities (Sodian & Frith, 1992). Furthermore, the relatively high egocentrism, low self-reflectivity, and impaired mindreading in Narcissistic Personality Disorder (NPD) shows a link between impaired ability to accurately label emotion within oneself and by extension extrapolate this information to others (Dimaggio, Semerari, Carcione, Nicolō, & Procacci, 2007). Persons with Anorexia Nervosa also have been shown to have higher scores of alexithymia, which may partly explain their deficits in social functioning (Taylor et al., 1996). Individuals with schizophrenia show neural deficits in offline and online mentalizing tasks (Russell et al., 2000, Brunet et al., 2003, Andreasen et al., 2008, Walter et al., 2009; Das et al., 2009). Furthermore, there is evidence showing an over-interpretation of the mental states of others, named "over-mentalizing" may underpin social anxiety disorder (Hezel and McNally, 2014). It is unclear, however, whether individuals can use the information gleaned from others in order to better understand their own personal feelings. It is therefore imperative to define the structure of components of mentalizing and study the heterogeneity of mentalizing deficits among the above mentioned clinical samples.

9. Concluding remarks

Our IMT framework integrates metacognition and mentalizing, and presents the new concepts of personal second-order mentalizing, and collective mentalizing. We believe that the four proposed elements of interactive mentalizing are important because it presents a theoretical framework for how social interactions impact the perception of ourselves, others, and our selves via others (e.g., reputation management). Our theory also provides a basis to empirically dissect social conditions that evoke overlapping and distinct mentalizing circuits. In turn, this will advance our understanding of the dynamic nature of social interaction and

insights into what circuits are most affected in autism, schizophrenia and other psychiatric disorders.

References

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. Journal of Cognitive Neuroscience, 16(10), 1773-1784. doi:Doi 10.1162/0898929042947928

Apps, M. A. J., Lesage, E. & Ramnani, N. Vicarious reinforcement learning signals when instructing others. J. Neurosci. 35, 2904–2913 (2015).

Andreasen, N. C., Calage, C. A., & O'Leary, D. S. (2008). Theory of mind and schizophrenia: a positron emission tomography study of medication-free patients. *Schizophrenia bulletin*, *34*(4), 708-719.

Atique, B., Erb, M., Gharabaghi, A., Grodd, W., & Anders, S. (2011). Task-specific activity and connectivity within the mentalizing network during emotion and intention mentalizing. Neuroimage, 55(4), 1899-1911

Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. Philos Trans R Soc Lond B Biol Sci, 367(1594), 1350-1365. doi:10.1098/rstb.2011.0420

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. Science, 329(5995), 1081-1085. doi:10.1126/science.1185718

Bandura, A., Ross, D. & Ross, S. A. Vicarious reinforcement and imitative learning. *J. Abnorm. Soc. Psychol.* **67**, 601–607 (1963).

Banerjee, R. (2002). Audience effects on self−presentation in childhood. Social Development, 11(4), 487-507.

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. Proc Natl Acad Sci U S A, 115(23), 6082-6087. doi:10.1073/pnas.1800795115

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition, 21(1), 37-46.

Baron−Cohen, S., Leslie, A. M., & Frith, U. (1986). Mechanical, behavioural and intentional understanding of picture stories in autistic children. British Journal of developmental psychology, 4(2), 113-125.

Berecz, H., Tényi, T., & Herold, R. (2016). Theory of mind in depressive disorders: A review of the literature. Psychopathology, 49(3), 125-134.

Berns, G. S., Capra, C. M., Moore, S., & Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. Neuroimage, 49(3), 2687-2696. doi: 10.1016/j.neuroimage.2009.10.070

Berns, G. S., Chappelow, J., Zink, C. F., Pagnoni, G., Martin-Skurski, M. E., & Richards, J. (2005). Neurobiological correlates of social conformity and independence during mental rotation. Biological Psychiatry, 58(3), 245-253. doi:10.1016/j.biopsych.2005.04.012

Biele, G., Rieskamp, J., Krugel, L. K., & Heekeren, H. R. (2011). The neural basis of following advice. PLoS-Biology, 9(6), 1208.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. The Journal of Economic Perspectives, 151-170.

Blaffer Hrdy, S. (2009). Mothers and others: the evolutionary origins of mutual understanding. Cambridge: Belknap.

Blakemore, S. J. (2008). The social brain in adolescence. Nature Reviews Neuroscience, 9(4), 267-277. doi:10.1038/nrn2353

Bögels, S., Barr, D. J., Garrod, S., & Kessler, K. (2014). Conversational interaction in the scanner: mentalizing during language processing as revealed by MEG. Cerebral Cortex, 25(9), 3219-3234.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. Nature, 489(7415), 295-298.

Brüne, M. (2005). "Theory of mind" in schizophrenia: a review of the literature. Schizophrenia Bulletin, 31(1), 21-42.

Brunet, E., Sarfati, Y., Hardy-Baylé, M. C., & Decety, J. (2003). Abnormalities of brain function during a nonverbal theory of mind task in schizophrenia. Neuropsychologia, 41(12), 1574-1582.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. Trends in cognitive sciences, 11(2), 49-57.

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. Proceedings of the National Academy of Sciences, 107(32), 14431-14436.

Burnkrant, R. E., & Cousineau, A. (1975). Informational and normative social influence in buyer behavior. Journal of Consumer research, 206-215.

Camerer, C. and Mobbs, D. (2017). Comparing cognitive and neural processes during hypothetical and real choices. Trends in Cognitive Science. 21, 1, 46-56.

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. Neuroimage, 12(3), 314-325.

Charness, G., Rustichini, A., & van de Ven, J. (2018). Self-confidence and strategic behavior. Experimental Economics, 21(1), 72-98. doi:10.1007/s10683-017-9526-3

Choi-Kain, L. W., & Gunderson, J. G. (2008). Mentalization: Ontogeny, assessment, and application in the treatment of borderline personality disorder. American Journal of Psychiatry, 165(9), 1127-1135.

Chua, E. F., Schacter, D. L., Rand-Giovannetti, E., & Sperling, R. A. (2006). Understanding metamemory: neural correlates of the cognitive process and subjective level of confidence in recognition memory. *Neuroimage, 29*(4), 1150-1160.

Chua, E. F., Schacter, D. L., & Sperling, R. A. (2009). Neural basis for recognition confidence in younger and older adults. *Psychol Aging, 24*(1), 139-153.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. Annu. Rev. Psychol., 55, 591-621.

Cobb, M. D., & Kuklinski, J. H. (1997). Changing minds: Political arguments and political persuasion. American Journal of Political Science, 88-121.

Das, P., Lagopoulos, J., Coulston, C. M., Henderson, A. F., & Malhi, G. S. (2012). Mentalizing impairment in schizophrenia: a functional MRI study. Schizophrenia research, 134(2-3), 158-164.

Devaine, M., Hollard, G., & Daunizeau, J. (2014). The social Bayesian brain: does mentalizing make a difference when we learn?. PLoS computational biology, 10(12), e1003992.

Dimaggio, G., Lysaker, P. H., Carcione, A., Nicolo, G., & Semerari, A. (2008). Know yourself and you shall know the other… to a certain extent: multiple paths of influence of self-reflection on mindreading. Consciousness and Cognition, 17(3), 778-789.

Dimaggio, G., Semerari, A., Carcione, A., Nicolō, G., & Procacci, M. (2007). Psychotherapy of personality disorders: Metacognition, states of mind and interpersonal cycles: Routledge.

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In Advances in experimental social psychology (Vol. 44, pp. 247-296). Academic Press.

Echterhoff, G., Higgins, E. T., & Levine, J. M. (2009). Shared Reality Experiencing Commonality With Others' Inner States About the World. *Perspectives on Psychological Science, 4*(5), 496-521. doi:10.1111/j.1745-6924.2009.01161.x

Edelson, M., Sharot, T., Dolan, R. J., & Dudai, Y. (2011). Following the Crowd: Brain Substrates of Long-Term Memory Conformity. Science, 333(6038), 108-111. doi:10.1126/science.1203557

Efklides, A., & Misailidi, P. (2010). Introduction: The present and the future in metacognition Trends and prospects in metacognition research (pp. 1-18): Springer.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American psychologist, 34(10), 906.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. Philosophical Transactions of the Royal Society B-Biological Sciences, 367(1594), 1338-1349. doi:10.1098/rstb.2011.0417

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. Frontiers in Human Neuroscience, 8, 443.

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. Science, 329(5998), 1541-1543.

Fonagy, P., & Luyten, P. (2009). A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. J Development psychopathology, 21(4), 1355-1381.

Frith, C., & Frith, U. (2005). Theory of mind. Current biology, 15(17), R644-R645.

Frith, C. D. (2012). The role of metacognition in human social interactions. Philosophical Transactions of the Royal Society B-Biological Sciences, 367(1599), 2213-2223. doi: 10.1098/rstb.2012.0123

Frith, C. D., & Frith, U. (1999). Interacting minds--a biological basis. Science, 286(5445), 1692-1695.

Frith, U., & Frith, C. (2001). The biological basis of social interaction. Current Directions in Psychological Science, 10(5), 151-155. doi:Doi 10.1111/1467-8721.00137

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. Trends in Cognitive Sciences, 7(2), 77-83.

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. Neuroimage, 16(3), 814-821.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. Trends Cogn Sci, 2(12), 493-501.

Gao, T., Scholl, B. J., & McCarthy, G. (2012). Dissociating the detection of intentionality from animacy in the right posterior superior temporal sulcus. J Neurosci, 32(41), 14276-14280. doi:10.1523/JNEUROSCI.0562-12.2012

Gilbert, S. J., Spengler, S., Simons, J. S., Steele, J. D., Lawrie, S. M., Frith, C. D., & Burgess, P. W. (2006). Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. Journal of cognitive neuroscience, 18(6), 932-948.

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: a comparison of theory of mind tasks. Journal of cognitive neuroscience, 19(11), 1803-1814.

Górska, D., & Marszał, M. (2014). Mentalization and theory of mind in borderline personality organization: exploring the differences between affective and cognitive aspects of social cognition in emotional pathology. Psychiatria polska, 48(3), 503-513.

Granhag, P. A., & Hartwig, M. (2008). A new theoretical perspective on deception detection: On the psychology of instrumental mind-reading. Psychology, Crime Law and Human Behavior, 14(3), 189-200.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. Science, 315(5812), 619-619.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci U S A, 105(18), 6741-6746. doi:10.1073/pnas.0711099105

Hagan, C. C. (2009). Mechanisms underlying the integration of seen and heard human emotion (Doctoral dissertation, The University of York).

Hagan, C. C., Woods, W., Johnson, S., Green, G. G., & Young, A. W. (2013). Involvement of right STS in audio-visual integration for affective speech demonstrated using MEG. PloS one, 8(8), e70648.

Hardin, C. D., & Conley, T. D. (2001). A relational approach to cognition: Shared experience and relationship affirmation in social cognition In Moskowitz GB, editor.(Ed.), Cognitive social psychology: The Princeton Symposium on the legacy and future of social cognition (pp. 3–17).

Hardin, C. D., & Higgins, E. T. (1996). Shared reality: How social verification makes the subjective objective.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. Biol Psychiatry, 51(1), 59-67.

Hezel, D. M., & McNally, R. J. (2014). Theory of mind impairments in social anxiety disorder. Behavior Therapy, 45, 530–540.

Hill, M. R., Boorman, E. D. & Fried, I. (2016). Observational learning computations in neurons of the human anterior cingulate cortex. *Nat. Commun.* 7, 12722.

Hongkeun, K., & Roberto, C. (2007). Trusting our memories: dissociating the neural correlates of confidence in veridical versus illusory memories. *Journal of Neuroscience, 27*(45), 12190-12197.

Hutto, D. D., & Ratcliffe, M. (2007). Folk psychology re-assessed: Springer.

Izuma, K., & Adolphs, R. (2013). Social Manipulation of Preference in the Human Brain. Neuron, 78(3), 563-573. doi:10.1016/j.neuron.2013.03.023

Izuma, K., Matsumoto, M., Murayama, K., Samejima, K., Sadato, N., & Matsumoto, K. (2010). Neural correlates of cognitive dissonance and choice-induced preference change. Proceedings of the National Academy of Sciences of the United States of America, 107(51), 22014-22019. doi:10.1073/pnas.1011879108

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. Science, 342(6156), 377-380.

Kim, H., & Cabeza, R. (2009). Common and specific brain regions in high- versus low-confidence recognition memory. *Brain Research, 1282*, 103-113.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. Psychological Review, 103(3), 490-517.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol, 77(6), 1121-1134. doi:Doi 10.1037/0022-3514.77.6.1121

Kuhn, D. (2000). Metacognitive development. Current Directions in Psychological Science, 9(5), 178-181.

Lee, S. M., Gao, T., & McCarthy, G. (2014). Attributing intentions to random motion engages the posterior superior temporal sulcus. Soc Cogn Affect Neurosci, 9(1), 81-87. doi:10.1093/scan/nss110

Liszkowski, U. (2018). Emergence of shared reference and shared minds in infancy. *Current opinion in psychology*, *23*, 26-29.

Liotti, G., & Gilbert, P. (2011). Mentalizing, motivation, and social mentalities: Theoretical considerations and implications for psychotherapy. Psychology and Psychotherapy-Theory Research and Practice, 84(1), 9-25. doi:10.1348/147608310x520094

Lockwood, P. L., Apps, M. A., Roiser, J. P. & Viding, E. (2015).Encoding of vicarious reward prediction in anterior cingulate cortex and relationship with trait empathy. *J. Neurosci. 35*, 13720–13727

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Baron-Cohen, S., & Consortium, M. A. (2011). Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. Neuroimage, 56(3), 1832-1838. doi:10.1016/j.neuroimage.2011.02.067

Mahmoodi, A., Bang, D., Ahmadabadi, M. N., & Bahrami, B. (2013). Learning to Make Collective Decisions: The Impact of Confidence Escalation. Plos One, 8(12), e81195. doi:10.1371/journal.pone.0081195

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: meta-d′, response-specific meta-d′, and the unequal variance SDT model The cognitive neuroscience of metacognition (pp. 25-66): Springer.

Marshall, J. A. R., Brown, G., & Radford, A. N. (2017). Individual Confidence-Weighting and Group Decision-Making. Trends in Ecology & Evolution, 32(9), 636-645. doi:https://doi.org/10.1016/j.tree.2017.06.004

McDonald, S., & Flanagan, S. (2004). Social perception deficits after traumatic brain injury: Interaction between emotion recognition, mentalizing ability, and social communication. Neuropsychology, 18(3), 572-579. doi:10.1037/0894-4105.18.3.572

Metcalfe, J., & Shimamura, A. P. (1994). Metacognition: Knowing about knowing: MIT press.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. Journal of Cognitive Neuroscience, 17(8), 1306-1315. doi:Doi 10.1162/0898929055002418

Molenberghs, P., Trautwein, F.-M., Böckler, A., Singer, T., & Kanske, P. (2016). Neural correlates of metacognitive ability and of feeling confident: a large-scale fMRI study. Social Cognitive and Affective Neuroscience, 11(12), 1942-1951. doi:10.1093/scan/nsw093

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings Psychology of learning and motivation (Vol. 26, pp. 125-173): Elsevier.

Nook, E. C., & Zaki, J. (2015). Social Norms Shift Behavioral and Neural Responses to Foods. Journal of Cognitive Neuroscience, 27(7), 1412-1426. doi:10.1162/jocn_a_00795

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: the neural systems of social fear transmission. Soc Cogn Affect Neurosci, 2(1), 3-11. doi: 10.1093/scan/nsm005

Olson, I. R., Ploaker, A., & Ezzyat, Y. (2007). The Enigmatic temporal pole: a review of findings on social and emotional processing. Brain, 130, 1718-1731. doi:10.1093/brain/awm052

Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. J Cogn Neurosci, 16(10), 1706-1716. doi:10.1162/0898929042947900

Petty, R., Briñol, P., & L Tormala, Z. (2002). Thought confidence as a determinant of persuasion: The self-validation hypothesis (Vol. 82).

Petty, R. E., Briñol, P., Tormala, Z. L., & Wegener, D. T. (2007). The role of meta-cognition in social judgment. Social psychology: Handbook of basic principles, 2, 254-284.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral brain sciences, 1(4), 515-526.

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. Journal of Experimental Social Psychology, 43(4), 565-578. doi:https://doi.org/10.1016/j.jesp.2006.05.011.

Qi, S., Footer, O., Camerer, C., and Mobbs, D. (2018). A collaborator's reputation can bias decisions and anxiety under uncertainty. Journal of Neuroscience. 38 (9), 2262-2269.

Redcay, E. (2008). The superior temporal sulcus performs a common function for social and speech perception: implications for the emergence of autism. Neurosci Biobehav Rev, 32(1), 123-142. doi:10.1016/j.neubiorev.2007.06.004

Redcay, E., & Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. Nature Reviews Neuroscience. doi:10.1038/s41583-019-0179-4

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical beliefs. Current Biology, 28(24), 4014-4021.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". Neuroimage, 19(4), 1835-1842.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. Neuropsychologia, 42(11), 1435-1446.

Schultz, J., Imamizu, H., Kawato, M., & Frith, C. D. (2004). Activation of the human superior temporal gyrus during observation of goal attribution by intentional objects. J Cogn Neurosci, 16(10), 1695-1705. doi:10.1162/0898929042947874

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. Proceedings of the National Academy of Sciences of the United States of America, 104(18), 7361-7366. doi:10.1073/pnas.0702077104

Russell, T. A., Rubia, K., Bullmore, E. T., Soni, W., Suckling, J., Brammer, M. J., ... & Sharma, T. (2000). Exploring the social brain in schizophrenia: left prefrontal underactivation during mental state attribution. American journal of psychiatry, 157(12), 2040-2042.

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. Journal of retailing, 80(2), 159-169.

Silston, B., Bassett, D. S., & Mobbs, D. (2018). How Dynamic Brain Networks Tune Social Behavior in Real Time. Current Directions in Psychological Science, 27(6), 413-421.

Slaughter, V., Peterson, C. C., & Moore, C. (2013). I can talk you into it: Theory of mind and persuasion behavior in young children. Developmental Psychology, 49(2), 227.

Sodian, B., & Frith, U. (1992). Deception and sabotage in autistic, retarded and normal children. J Child Psychol Psychiatry, 33(3), 591-605.

Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. Journal of cognitive neuroscience, 21(3), 489-510.

Stallen, M., & Sanfey, A. G. (2015). The neuroscience of social conformity: implications for fundamental and applied research. Front Neurosci, 9, 337. doi:10.3389/fnins.2015.00337

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. Journal of Personality and Social Psychology, 48(6), 1467.

Taylor, S. E., & Brown, J. D. (1988). Illusion and Well-Being - a Social Psychological Perspective on Mental-Health. Psychological Bulletin, 103(2), 193-210. doi:Doi 10.1037/0033-2909.103.2.193

Taylor, G. J., Parker, J. D., Bagby, R. M., & Bourke, M. P. (1996). Relationships between alexithymia and psychological characteristics associated with eating disorders. *Journal of Psychosomatic research*, *41*(6), 561-568.

Thorley, C. (2013). Memory conformity and suggestibility. Psychology, Crime & Law, 19(7), 565-575. doi:10.1080/1068316X.2011.648637

Thornton, M. A., Weaverdyck, M. E., Mildner, J. N., & Tamir, D. I. (2019). People represent their own mental states more distinctly than those of others. Nature communications, 10(1), 2117.

Toelch, U., & Dolan, R. J. (2015). Informational and Normative Influences in Conformity from a Neurocomputational Perspective. Trends Cogn Sci, 19(10), 579-589. doi:10.1016/j.tics.2015.07.007

Tormala, Z. L., & Petty, R. E. (2004). Source Credibility and Attitude Certainty: A Metacognitive Analysis of Resistance to Persuasion. Journal of Consumer Psychology, 14(4), 427-442. doi:https://doi.org/10.1207/s15327663jcp1404_11

Tuomela, R. (2006). Joint Intention, We-Mode and I-Mode. Midwest Studies In Philosophy, 30(1), 35-58. doi:10.1111/j.1475-4975.2006.00127.x

Vanhuyck, J. B., Battalio, R. C., & Beil, R. O. (1990). Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure. American Economic Review, 80(1), 234-248.

Walter, H., Ciaramidaro, A., Adenzato, M., Vasic, N., Ardito, R. B., Erk, S., & Bara, B. G. (2009). Dysfunction of the social brain in schizophrenia is modulated by intention type: an fMRI study. Social cognitive and affective neuroscience, 4(2), 166-176.

Wang, Z. L. (2015). Theory of mind and children's understanding of teaching and learning during early childhood. Cogent Education, 2(1). doi:Artn 1011973 10.1080/2331186x.2015.1011973

Wheeler, R., Allan, K., Tsivilis, D., Martin, D., & Gabbert, F. (2013). Explicit Mentalizing Mechanisms and Their Adaptive Role in Memory Conformity. Plos One, 8(4), e62106. doi:10.1371/journal.pone.0062106

Wood, W. (2000). Attitude change: Persuasion and social influence. Annual review of psychology, 51(1), 539-570.

Wu, H., Fung, B. J., & Mobbs, D. (2019, July 31). Mentalizing during social interaction: the development and validation of the interactive mentalizing questionnaire. https://doi.org/10.31234/osf.io/g2zm8

Wu, H., Luo, Y., & Feng, C. (2016). Neural signatures of social conformity: A coordinate-based activation likelihood estimation meta-analysis of functional brain imaging studies. Neurosci Biobehav Rev, 71, 101-111. doi:10.1016/j.neubiorev.2016.08.038

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. PLoS computational biology, 4(12), e1000254.

Zaki, J., & Ochsner, K. N. (2012). The neuroscience of empathy: progress, pitfalls and promise. Nature neuroscience, 15(5), 675.

Zunshine, L. (2006). Why we read fiction: Theory of mind and the novel: Ohio State University Press.